Original Article

Accuracy and reliability of automated landmark identification and cephalometric measurements on cone beam computed tomography using Invivo software

Young-Eun Jung[a]; Heeyeon Suh[b]; Joorok Park[c]; Heesoo Oh[d]

ABSTRACT

Objectives: To evaluate the accuracy and reliability of an automated landmark identification (ALI) system and the impact of ALI errors on cephalometric measurements on cone-beam computed tomography (CBCT) images.

Materials and Methods: Thirty-one landmarks were identified on 76 CBCT images using Invivo7 software (Anatomage, San Jose, Calif). Ground truth was established by averaging landmark coordinates from two calibrated human examiners. The accuracy of the ALI system was assessed by the mean absolute error (MAE, mm) across coordinate axes, the mean error distance (mm), and the successful detection rate (SDR) for each landmark. Interexaminer reliability between the ALI and manual landmark location was evaluated. Eighteen cephalometric measurements were computed from 25 landmarks. Accuracy of measurements from the ALI system was assessed with the MAE and successful measurement rates (SMR).

Results: The ALI system closely matched human examiners in landmark identification, with an average MAE of 0.94 $\pm$ 0.99 mm. Across all three coordinate axes, 87% of the landmarks had $<$2 mm MAE. ALI average MAE for conventional linear and angular cephalometric measurements were 1.35 $\pm$ 1.33 mm and 0.89 $\pm$ 0.89 degrees, respectively. Only one measurement, Intercondylar Width, showed MAE $>$3 mm.

Conclusions: The ALI system showed clinically acceptable accuracy and reliability for the majority of cephalometric landmarks and measurements. Clinicians are advised to critically evaluate ALI landmarks with substantial errors, to fully utilize the capabilities of commercial software effectively. (*Angle Orthod*. 2025;95:362–370.)

KEY WORDS: Automated; Landmark identification; Cephalometric analysis; CBCT

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-22 via free access

INTRODUCTION

Cephalometric analysis provides a quantitative method to assess skeletal and dentoalveolar morphology, which are crucial for understanding dentofacial discrepancies and growth.[1] Cone-beam computed tomography (CBCT) technology has enabled three-dimensional (3D) analysis, offering more precise anatomical representation.[2] A key advantage of 3D analysis is the elimination of bilateral structure superimposition and distortion present in 2D imaging, improving accuracy.[3]

Despite the benefits offered by CBCT, challenges remain, especially in the time-consuming and less familiar process of landmark identification on 3D images. Computational advancements have propelled machine learning applications, including automated landmark identification (ALI), facilitating analysis in both two-dimensional (2D)

[a] Clinical Assistant Professor, Department of Orthodontics, New York University, College of Dentistry, New York, NY, USA.
[b] Assistant Professor, Department of Orthodontics, Arthur A. Dugoni School of Dentistry, University of the Pacific, San Francisco, California, USA.
[c] Associate Professor, Department of Orthodontics, Arthur A. Dugoni School of Dentistry, University of the Pacific, San Francisco, California, USA.
[d] Professor and Chair, Department of Orthodontics, Arthur A. Dugoni School of Dentistry, University of the Pacific, San Francisco, California, USA.
Corresponding author: Dr Heeyeon Suh, Department of Orthodontics, Arthur A. Dugoni School of Dentistry, University of the Pacific, San Francisco, California 94103, USA
(e-mail: hsuh1@pacific.edu)

Accepted: March 10, 2025. Submitted: December 23, 2024.
Published Online: April 10, 2025
© 2025 by The EH Angle Education and Research Foundation, Inc.

and 3D imaging.[4–16] Artificial intelligence (AI) integration aims to lessen clinician workload and expedite the process.[17] ALI software reduces identification time to 1–2 minutes compared to the 15 minutes needed for manual CBCT identification by an experienced operator.[1,2,4,7,8]

Recent research has shown that deep learning methods accurately detect landmarks.[1,8,9] Promising algorithms developed over the past two years have improved ALI accuracy on 3D images.[14–16] Although previous studies demonstrated promising results, the impact on cephalometric analysis remains to be evaluated thoroughly.[13,18,19] This study had two main objectives: to assess the accuracy and reliability of automated 3D CBCT landmark identification using the widely used commercial software, Invivo (Anatomage, San Jose, Calif), and to evaluate the impact of ALI errors on commonly used cephalometric measurements.

MATERIALS AND METHODS

This study was approved by the institutional review board of the University of the Pacific (#2021-95). The study sample was gathered retrospectively from University of Pacific Department of Orthodontics. The inclusion criteria were: (1) CBCT scans with a voxel size $\leq$0.3 mm^3 and a field of view at least 16 $\times$ 13 cm, and (2) the presence of all permanent teeth. Participants of all ages, genders, and with various skeletal conditions were included. Exclusion criteria were: (1) restorative work causing significant scatter and (2) the presence of craniofacial abnormalities, syndromes, or cleft lip and palate.

The sample consisted of 76 CBCT volumes, acquired using an Imaging Science International CBCT scanner (Hatfield, PA). These volumes were imported in Digital Imaging and Communications in Medicine (DICOM) format into the Invivo7 software (Anatomage, San Jose, Calif). The software included multiplanar sectional slices in axial, sagittal, and coronal views to aid in identifying the landmarks (Figure 1). CBCT images were traced independently by two calibrated human examiners (YJ and HS) and the ALI system. A total of 31 landmarks were utilized, comprising 17 skeletal, eight soft tissue, and six dental landmarks (Table 1). The resulting x, y, and z coordinates were exported to a Microsoft Excel (Microsoft Corp., Redmond, Wash) file. The ground truth was established by calculating the mean x, y, and z coordinates by two examiners for each landmark.

Cephalometric measurements were calculated using coordinate values reoriented to a standard anatomical frame of reference (AFOR). The axial plane was formed by right Porion, left Porion, and right Orbitale. The sagittal plane was constructed perpendicular to the axial plane and contained Nasion and Basion. The coronal plane was made perpendicular to the other two planes and passed through Nasion. Eighteen cephalometric measurements, including eight angular and 10 linear measurements, were computed. All angular and eight linear measurements were calculated by projecting the line segments onto the midsagittal plane. Transverse width measurements, intercondylar width and mandibular width, were projected onto the coronal plane. Distance and angle calculations used standard mathematical formulas.

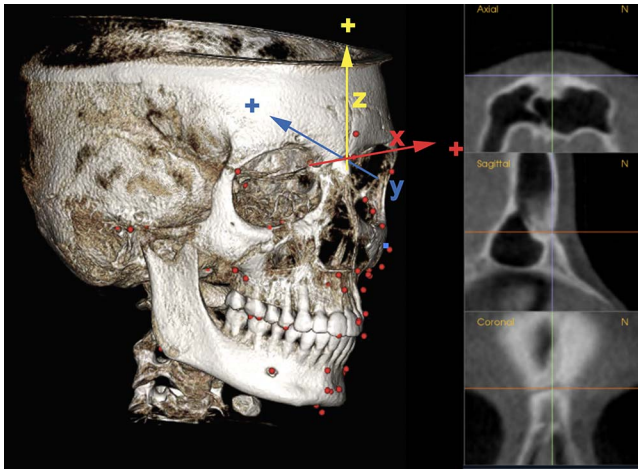

Figure 1. Three-dimensional landmark identification on CBCT. Positive signs on the x, y, and z axes denote left, posterior, and upward directions, respectively, while negative signs indicate right, anterior, and downward.

To evaluate the accuracy of ALI, the mean absolute error (MAE, mm) in the x, y, and z coordinates between the ground truth and ALI were calculated and the error distance was calculated with a 3D Euclidian distance formula.[3,10] A successful detection rate (SDR) was also calculated. The SDR represents the percentage of images in which the landmark was located within a precision range.[3,11] Clinically, an ALI system is considered accurate if the variance from the ground truth is less than 2 mm, and acceptable if under 4 mm.[5,14,20] To detect subtle changes during treatment or growth, this study implemented stricter criteria of $\leq$1 mm, 1.5 mm, 2 mm, and 3 mm. Successful measurement rate (SMR) was also calculated with the same criteria as the SDR. To evaluate intraexaminer reliability of the ALI system, all 76 images were subjected to two rounds of testing at an interval of 1 week. The interexaminer reliability between ALI and manual identification, as well as the reliability between two calibrated human examiners, was evaluated.

Statistical Analysis

Basic descriptive statistics including the mean, standard deviation (SD), and percentages were computed.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-22 via free access

Table 1. Definitions Used by Human Examiners for 3D Landmark Identification

Category	Landmark	Definition
Skeletal	Nasion	The most anterior and median point along the frontonasal suture
	Sella	The geometric center of the sella turcica
	Basion	The most inferior point on the anterior border of the foramen magnum on the midsagittal plane
	Orbitale (left and right)	The lowest point on the infraorbital margin
	Porion (left and right)	Most lateral and superior point located on the external auditory meatus
	ANS	Anterior nasal spine, most anterior point on the maxilla
	PNS	Posterior nasal spine, the most posterior point on the sagittal plane of the bony hard palate on the maxillary midline
	A-point	A midline point in the deepest concavity along the anterior contour of the maxilla
	B-point	A midline point in the deepest concavity on the anterior contour on the lower alveolar arch between the chin and the mandibular alveolar process
	Pogonion	Most anterior point on the symphysis of the mandible on the mandibular midline
	Menton	Lower most point on the mandibular symphysis on the mandibular midline
	Gonion (left and right)	A point on the bony contour of the gonial angle determined by bisecting the tangent angle and the lowest point on the coronal section
	Condylion (left and right)	The most superior point on the condyle of the mandible
Soft tissue	Soft tissue nasion	Most posterior point on the soft tissue profile between glabella and pronasale on the midsagittal plane
	Pronasale	The most anterior point on the nose tip
	Upper lip	Most prominent point of the upper lip on the upper lip midline
	Stomion superius	Most inferior point located on the upper lip in the middle of upper lip
	Stomion inferius	Most superior point located on the lower lip in the middle of lower lip
	Lower lip	Most prominent point of the lower lip on the lower lip midline
	Soft tissue B-point	The deepest point on concavity between lower lip and soft tissue pogonion on the mandibular midline
	Soft tissue pogonion	Most anterior point on the soft tissue chin on the mandibular midline
Dental	Upper incisal root	Root apex of the maxillary right central incisor
	Upper incisal crown	Middle of the incisal edge of the maxillary right central incisor
	Lower incisal root	Root apex of the mandibular right central incisor
	Lower incisal crown	Middle of the incisal edge of the mandibular right central incisor
	Upper molar cusp	The mesiobuccal cusp tip of the maxillary right first molar
	Lower molar cusp	The mesiobuccal cusp tip of the mandibular right first molar

Intraclass correlation coefficients (ICC) were used to evaluate reliability. To visualize and compare reliability, scattergrams with 95% confidence ellipses were generated, showing differences between ALI and human examiner 2, as well as differences between human examiners 1 and 2. Data were analyzed using Statistical Package for the Social Sciences (IBM Corp, Armonk, New York) and language R (Vienna, Austria).

RESULTS

The ALI system demonstrated perfect consistency, with an ICC of 1 when the same CBCT images were processed twice. The interexaminer reliability for landmark location between the two calibrated human examiners was excellent, with ICC ranging from 0.9 to 1 except for the x-coordinate of left Orbitale with 0.7. The interexaminer reliability between ALI and a human examiner (human examiner 2) ranged from 0.9 to 1 for the y and z coordinates, and from 0.5 to 1 for the x coordinates. In general, the landmarks which exhibited large differences between human examiners also showed considerable differences between the human and ALI (Figure 2). However, there were exceptions such as Porion, Condylion, and the maxillary first molar cusp: the x-coordinates for Porion demonstrated the lowest reliability, with ICC values ranging from 0.5 to 0.6, compared to an ICC of 0.9 between two calibrated human examiners. Both x-coordinates of Condylion and the maxillary first molar cusp exhibited lower reliability (ICC $= 0.7$) compared to the ICC >0.9 observed between human examiners.

The overall accuracy of the ALI system in comparison to the established ground truth is presented in Table 2. ALI achieved remarkable accuracy, with a mean absolute error (MAE) <2 mm for all landmarks, except for Porion, Gonion, and Stomion Superius. The MAE for the 31 landmarks was 1.35 mm on the x-axis, 0.72 mm on the y-axis, and 0.74 mm on the z-axis, resulting in an overall MAE of 0.94 mm. The mean error distance between the ALI system and ground truth was 1.99 ± 1.26 mm, with 81% of the landmarks showing a SDR within a 3 mm error distance margin. Notably, Nasion and Sella landmarks showed exceptional accuracy, achieving SDRs of 91% and 92% within a 2 mm error distance range (Table 3). High accuracy was also observed for the upper and lower incisor edges, attaining 100% and 99% SDRs within 2 mm, whereas the lower incisor root

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-22 via free access

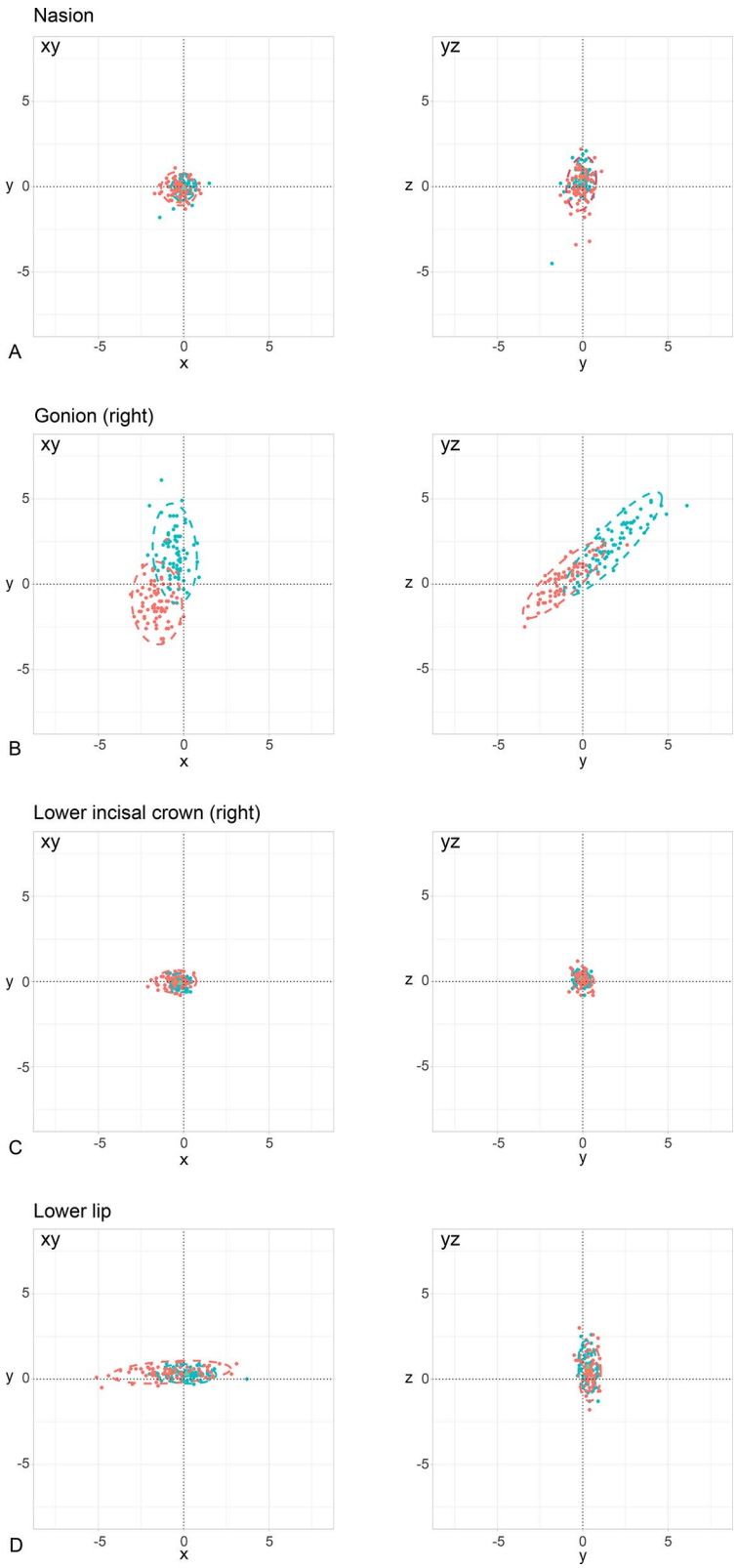

Figure 2. Scatterplots of (ALI coordinates – human examiner 2 coordinates) (red), and (human examiner 1 coordinates – human examiner 2 coordinates) (blue), with 95% confidence ellipses in different planes of view. Larger ellipses indicate greater variability and, thus, lower reliability. For bilateral landmarks, the right-side landmark is presented. Positive values on the x, y, and z axes indicate that the ALI (red) or human examiner 1 (blue) placed landmarks medially, posteriorly, and superiorly compared to the locations identified by human examiner 2. (A), Nasion; (B), Gonion; (C), Lower incisal crown; (D), Lower lip.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-22 via free access

Table 2. Mean Absolute Error at x, y, and z Coordinates for Each Landmark and Mean Error Distance for Each Landmark. Errors Calculated Using the Average of Human Examiner Coordinates as the Reference Standard

	Mean Absolute Error (mm)						Mean Error Distance (mm)	
	x		y		z			
Landmark	mean	SD	mean	SD	mean	SD	Mean	SD
Skeletal								
Nasion	0.43	0.38	0.33	0.28	0.75	0.69	1.05	0.66
Sella	0.54	0.44	0.54	0.43	0.58	0.45	1.10	0.53
Basion	1.13	1.47	0.48	0.51	0.49	0.45	1.48	1.47
Orbitale (left)	1.47	1.22	0.96	0.70	0.29	0.21	1.90	1.26
Orbitale (right)	1.27	0.91	0.76	0.63	0.33	0.29	1.64	0.96
Porion (left)	3.06	1.29	0.81	0.98	0.60	0.48	3.38	1.36
Porion (right)	3.57	1.86	0.65	0.56	0.55	0.49	3.76	1.83
ANS	0.68	0.52	0.87	0.82	0.65	0.55	1.49	0.81
PNS	0.60	0.51	0.92	1.00	0.67	0.83	1.51	1.16
A-point	0.94	0.77	0.27	0.22	1.02	0.81	1.60	0.85
B-point	1.57	1.22	0.23	0.19	1.44	1.06	2.36	1.28
Pogonion	1.79	1.29	0.30	0.26	1.16	0.90	2.40	1.17
Menton	1.84	1.34	0.58	0.50	0.37	0.36	2.11	1.26
Gonion (left)	1.48	0.57	2.13	0.85	1.36	0.92	3.09	0.95
Gonion (right)	1.30	0.61	2.02	0.97	1.16	0.83	2.84	0.95
Condylion (left)	1.90	0.86	0.58	0.44	1.21	0.69	2.48	0.81
Condylion (right)	1.47	0.96	0.81	0.63	1.31	0.85	2.36	1.00
Soft tissue								
Soft tissue nasion	0.40	0.35	0.52	0.36	1.90	1.24	2.09	1.20
Pronasale	1.26	1.04	0.20	0.16	0.53	0.42	1.51	0.97
Upper lip	1.21	0.95	0.21	0.16	0.54	0.45	1.45	0.9
Stomion superius	1.30	0.99	2.30	1.18	0.40	0.33	2.87	1.17
Stomion inferius	1.36	1.05	1.95	1.08	0.37	0.30	2.63	1.10
Lower lip	1.50	1.10	0.31	0.22	0.63	0.52	1.79	1.03
Soft tissue B-point	1.55	1.17	0.26	0.25	0.49	0.50	1.78	1.11
Soft tissue pogonion	1.77	1.23	0.29	0.23	1.35	1.13	2.51	1.25
Dental								
Upper incisal root	1.01	0.60	0.48	0.40	0.59	0.43	1.41	0.56
Upper incisal crown	0.68	0.45	0.39	0.27	0.25	0.27	0.93	0.40
Lower incisal root	1.08	0.57	0.63	0.61	0.88	0.65	1.68	0.78
Lower incisal crown	0.48	0.42	0.22	0.17	0.28	0.21	0.68	0.38
Upper molar cusp	1.98	0.90	0.48	0.36	0.35	0.33	2.14	0.84
Lower molar cusp	1.35	0.83	0.73	0.43	0.45	0.34	1.73	0.73

apex showed a 70% SDR within 2 mm. (Table 3). Right and left Porion exhibited the least accuracy, with mean errors of 3.76 $\pm$ 1.83 mm and 3.38 $\pm$ 1.36 mm, respectively (Table 2).

Cephalometric measurement errors are presented in Table 4. The MAE for conventional linear and angular cephalometric measurements were 0.93 $\pm$ 0.92 mm and 0.89 $\pm$ 0.89 degrees, respectively. In the conventional linear measurement category, PFH had the greatest MAE of 1.52 $\pm$ 1.30 mm and 65% accuracy within 2 mm, whereas all other measures indicated MAE around 1 mm. Two transverse measurements evaluated in this study, intercondylar width and mandibular width, had MAE of 3.32 $\pm$ 1.61 mm and 2.74 $\pm$ 1.03 mm, respectively. The measurements with higher reliability, which exhibited smaller differences between human examiners, also demonstrated higher reliability between a human and the ALI system (Figure 3).

DISCUSSION

The ALI system reduced landmark identification time to 1–2 minutes, compared to 15 minutes required for manual location on CBCT by calibrated human operators. The ALI system nearly matched human accuracy, with 95% of landmarks within a 3 mm error. This surpassed previous research by Shahidi et al.[7] or was comparable to Ghowsi et al.[14] Another distinction in the current study was the demonstrated high accuracy of the ALI system in locating dental landmarks. Notably, the maxillary and mandibular incisal edges achieved SDRs of 100% and 99%, respectively, within a 2-mm error distance (Table 3).

Errors in landmark identification can stem from the difference in the operational landmark definition between the software and human examiners. For example, ALI places Condylion along the mandibular profile line, whereas human examiners typically mark it at the

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-22 via free access

Table 3. Successful Detection Rate (SDR) for −1, −1.5, −2, and −3 mm Range Criteria for Landmark Location

	SDR (%) - Mean Absolute Error												SDR (%) - Mean Error Distance			
	x				y				z							
Landmark	1 mm	1.5 mm	2 mm	3 mm	1 mm	1.5 mm	2 mm	3 mm	1 mm	1.5 mm	2 mm	3 mm	1 mm	1.5 mm	2 mm	3 mm
Skeletal																
Nasion	93.42	97.37	98.68	100.00	97.37	100.00	100.00	100.00	73.68	89.47	93.42	97.37	60.53	84.21	90.79	96.05
Sella	85.53	97.37	98.68	100.00	81.58	97.37	100.00	100.00	84.21	96.05	98.68	100.00	47.37	80.26	92.11	100.00
Basion	67.11	80.26	82.89	90.79	86.84	96.05	97.37	98.68	89.47	97.37	98.68	100.00	51.32	71.05	81.58	88.16
Orbitale (left)	44.74	61.84	69.74	90.79	55.26	78.95	90.79	98.68	100.00	100.00	100.00	100.00	23.68	52.63	64.47	82.89
Orbitale (right)	43.42	61.84	81.58	94.74	69.74	85.53	93.42	100.00	94.74	100.00	100.00	100.00	28.95	50.00	68.42	89.47
Porion (left)	5.26	10.53	18.42	48.68	72.37	88.16	94.74	97.37	80.26	96.05	97.37	100.00	2.63	6.58	15.79	43.42
Porion (right)	3.95	6.58	19.74	44.74	77.63	92.11	98.68	98.68	84.21	93.42	98.68	100.00	2.63	3.95	17.11	39.47
ANS	68.42	96.05	98.68	100.00	68.42	85.53	94.74	96.05	80.26	93.42	97.37	100.00	25.00	56.58	81.58	96.05
PNS	78.95	94.74	97.37	100.00	69.74	85.53	90.79	96.05	76.32	93.42	96.05	98.68	31.58	61.84	80.26	96.05
A-point	61.84	78.95	90.79	98.68	98.68	100.00	100.00	100.00	57.89	73.68	89.47	96.05	23.68	48.68	80.26	92.11
B-point	36.84	55.26	72.37	84.21	98.68	100.00	100.00	100.00	40.79	57.89	77.63	88.16	17.11	27.63	42.11	71.05
Pogonion	30.26	44.74	65.79	80.26	94.74	100.00	100.00	100.00	48.68	69.74	80.26	96.05	7.89	19.74	46.05	71.05
Menton	34.21	44.74	61.84	78.95	81.58	93.42	98.68	100.00	94.74	97.37	98.68	100.00	23.68	36.84	56.58	77.63
Gonion (left)	19.74	50.00	81.58	98.68	10.53	21.05	46.05	81.58	39.47	55.26	75.00	96.05	0.00	1.32	11.84	55.26
Gonion (right)	36.84	63.16	82.89	98.68	15.79	28.95	47.37	82.89	42.11	69.74	82.89	98.68	2.63	7.89	17.11	57.89
Condylion (left)	15.79	35.53	52.63	93.42	81.58	94.74	98.68	100.00	39.47	65.79	86.84	98.68	1.32	15.79	27.63	71.05
Condylion (right)	35.53	56.58	68.42	96.05	68.42	86.84	96.05	98.68	38.16	60.53	78.95	93.42	6.58	15.79	38.16	77.63
Soft tissue																
Soft tissue nasion	92.11	97.37	100.00	100.00	89.47	100.00	100.00	100.00	25.00	46.05	57.89	78.95	21.05	42.11	48.68	76.32
Pronasale	46.05	63.16	82.89	92.11	98.68	98.68	98.68	98.68	85.53	93.42	98.68	98.68	31.58	56.58	77.63	92.11
Upper lip	46.05	64.47	78.95	94.74	100.00	100.00	100.00	100.00	84.21	94.74	98.68	100.00	34.21	59.21	76.32	90.79
Stomion superius	42.11	61.84	76.32	90.79	17.11	26.32	38.16	67.11	96.05	98.68	98.68	100.00	3.95	14.47	25.00	47.37
Stomion inferius	46.05	61.84	75.00	90.79	21.05	38.16	47.37	81.58	97.37	100.00	100.00	100.00	3.95	19.74	30.26	64.47
Lower lip	36.84	60.53	73.68	86.84	100.00	100.00	100.00	100.00	78.95	93.42	98.68	100.00	21.05	48.68	69.74	84.21
Soft tissue B-point	32.89	55.26	69.74	86.84	97.37	98.68	100.00	100.00	85.53	96.05	97.37	98.68	26.32	50.00	65.79	84.21
Soft tissue pogonion	30.26	46.05	59.21	85.53	98.68	100.00	100.00	100.00	42.11	65.79	78.95	92.11	1.32	26.32	40.79	72.37
Dental																
Upper incisal root	52.63	76.32	93.42	100.00	86.84	97.37	100.00	100.00	82.89	94.74	100.00	100.00	22.37	60.53	82.89	100.00
Upper incisal crown	81.58	93.42	100.00	100.00	96.05	100.00	100.00	100.00	97.37	98.68	100.00	100.00	61.84	89.47	100.00	100.00
Lower incisal root	44.74	78.95	96.05	98.68	85.53	92.11	93.42	98.68	59.21	85.53	94.74	97.37	21.05	42.11	69.74	92.11
Lower incisal crown	88.16	93.42	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	84.21	93.42	98.68	100.00
Upper molar cusp	13.16	26.32	51.32	92.11	93.42	98.68	100.00	100.00	97.37	98.68	98.68	100.00	7.89	19.74	47.37	92.11
Lower molar cusp	35.53	52.63	76.32	97.37	78.95	93.42	100.00	100.00	90.79	98.68	100.00	100.00	13.16	36.84	64.47	96.05

highest point of the condyle, usually medial to the ALI location. This discrepancy leads to consistent overestimation of intercondylion width by the ALI system (Figure 3). In 2D cephalometry, this is less critical since Condylion is used to measure mandibular and ramus lengths. In contrast, for 3D evaluations of transverse dimensions, discrepancies become significant. Significant transverse errors were also found for Porion, which

Table 4. Mean Absolute Error for Each Measurement and Successful Measurement Rate (SMR) Within −1, −1.5, −2, and −3 (Degrees or mm) Range Criteria. Errors Calculated Using the Average of Human Examiner Measurements as the Reference Standard

	Mean Absolute Error (°/mm)		SMR (%)			
Measurement	Mean	SD	1 (°/mm)	1.5 (°/mm)	2 (°/mm)	3 (°/mm)
Angular (°)						
SNA	0.85	0.69	65.79	82.89	93.42	98.68
SNB	0.78	0.57	67.11	86.84	96.05	100.00
ANB	0.30	0.23	98.68	100.00	100.00	100.00
FMA	0.80	0.66	75.00	84.21	94.74	100.00
U1SN	1.76	1.17	28.95	47.37	64.47	86.84
IMPA	1.36	1.22	46.05	65.79	72.37	89.47
Occlusal Plane Angle (OPA)	0.80	0.79	75.00	85.53	92.11	96.05
Facial angle	0.45	0.38	90.79	98.68	98.68	100.00
Linear (mm)						
Mx Length	1.10	0.83	51.32	71.05	85.53	97.37
Mn Length	1.06	0.91	59.21	75.00	86.84	97.37
Upper Facial Height (UFH)	1.08	0.82	55.26	73.68	88.16	94.74
Lower Facial Height (LFH)	0.92	0.88	61.84	82.89	90.79	97.37
Anterior Facial Height (AFH)	0.92	1.00	77.63	85.53	90.79	94.74
Posterior Facial Height (PFH)	1.52	1.30	44.74	60.53	64.47	81.58
Intercondylar Width	3.32	1.61	5.26	15.79	21.05	40.79
Mandibular Width	2.74	1.03	4.00	12.00	24.00	60.00
Upper Lip to E line	0.40	0.32	96.00	97.33	100.00	100.00
Lower Lip to E line	0.45	0.44	92.00	97.33	98.67	100.00

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-22 via free access

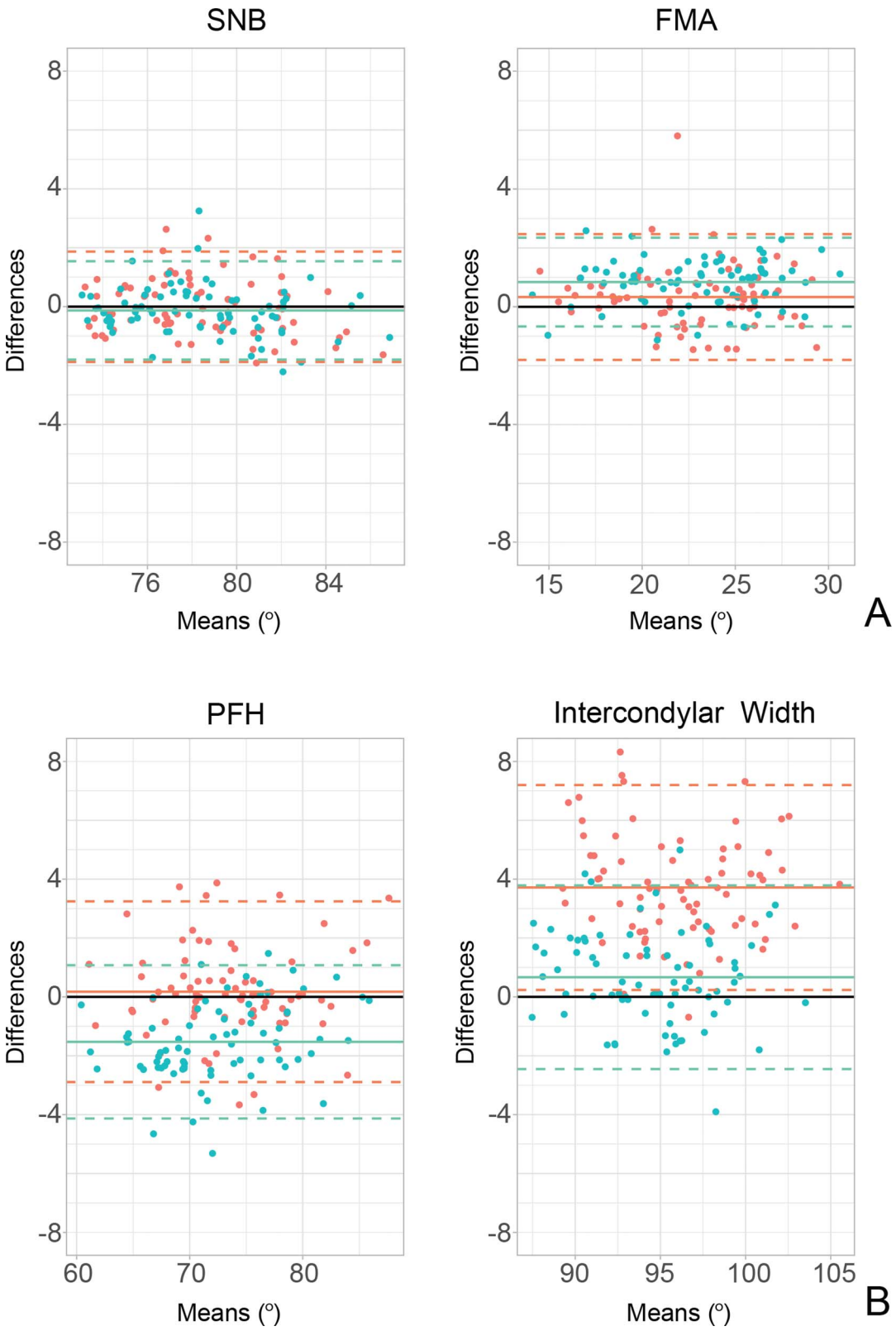

Figure 3. Bland-Altman plots presenting smaller variance (A) and larger variance (B) in measurements. Red represents (ALI – human examiner 2); Blue represents (human examiner 1 – human examiner 2). The solid line indicates the mean difference (bias); the dotted lines show the upper and lower limits of agreement. Wider limits of agreement indicate greater variability and, thus, lower reliability.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-22 via free access

ALI placed more medially on the temporal bone than human experts. However, these errors did not affect construction of the FH plane since anterior-posterior and vertical errors were smaller.

Landmark ambiguity contributes to identification errors. The definitions and anatomy allow for a degree of interpretation, leading to individual variance in landmark identification.[21,22] For instance, substantial errors across x, y, and z coordinates were observed for Gonion (Table 2). These errors may originate from the ALI training data from human experts, reflecting the challenge in defining Gonion on the broad curve of the mandible.[21,23,24] The difficulty in locating Gonion is illustrated in Figure 2B, which shows large differences between human examiners. Such discrepancies underscore the need for a clear consensus on definitions among human experts.[25]

Angular measurement errors tend to be larger when a line segment intersects the broader aspect of the error ellipse.[23,25] Despite substantial landmark location errors for Gonion, the mandibular plane angle (FMA) was accurate, as the larger error distribution dimension for Gonion aligned with the Gonion–Menton line (Figure 2B). The PFH measurement, a length measurement, exhibited greater error due to the larger error in Gonion landmark location. PFH showed larger differences both between human examiners and between the manual and ALI systems (Figure 3).

There is no definitive gold standard for identifying landmark positions in live patients (ie, directly identifying a site on the bone).[21] Although human examiners were calibrated, variations in landmark identification persist, leading to discrepancies in cephalometric measurements. Without a definitive ground truth, the intra- and interexaminer reliability patterns become essential for validating new methods.[26,27] This study showed that interexaminer reliability between the manual and ALI method was comparable to that between calibrated human examiners. Most variance arose from ambiguous landmark definitions, as shown by confidence ellipses (Figure 2) and limits of agreement (Figure 3). Although location of x-coordinates was less reliable between the manual and ALI methods, it did not significantly impact sagittal and vertical measurement errors. However, interpreting the intercondylar width requires caution, as the measurement was biased.

Improvements in ALI require refined 3D landmark definitions and expanded training datasets. Subjective interpretations by human experts lead to discrepancies, a challenge for ALI software development. Regarding measurements and analysis, if a landmark is to be used to evaluate a certain dimension, it should be shown to have relatively good consistency and precision.[21] For example, a more clear definition of Condylion point is necessary before intercondylar width can significantly contribute to transverse analysis.

The software used in this study detected mandibular midline landmarks on the midsagittal plane, which may result in large x-coordinate errors in patients with asymmetry (Figure 2D). These x-coordinate errors were observed in both the mandibular skeletal and soft tissue midline landmarks (Table 2). This indicated the need for caution when analyzing facial asymmetry and the importance of developing a more sophisticated detection algorithm for midline points. In addition, incorporating additional bilateral landmarks such as jugum, zygoma, first molars, and canines could enhance transverse analysis, which is a key advantage of 3D imaging.

This study provided an error range for each landmark in each dimension and the consequent impact on cephalometric measurement accuracy. Although ALI holds promise in orthodontic practice, it is imperative for clinicians to recognize the landmarks and measurements prone to substantial errors. A hybrid approach that combines ALI with human review for error-prone landmarks seems advisable for the time being. Additionally, developing an algorithm to highlight significant outliers for the clinician to review, and allowing manual adjustments, will be essential features for ALI software. The path forward requires more rigorous and standardized landmark definitions and the cautious use of less accurate landmarks and measurements.

CONCLUSIONS

- ALI achieved an SDR of 87% for all 31 landmarks within a 2-mm margin of error.
- Interexaminer reliability between a human and the ALI system was comparable to that between calibrated human examiners, with most variance arising from ambiguous definitions of the landmarks.
- The MAE of all measurements remained under 2 degrees or 2 mm for commonly used cephalometric measurements. The intercondylar width and mandibular width were not as accurate with MAE $>$2 mm.
- Clinicians must recognize potential inaccuracies in landmarks and measurements susceptible to significant errors, and more rigorous and standardized definitions are required to enhance ALI use in orthodontics.

REFERENCES

1. Bao H, Zhang K, Yu C, et al. Evaluating the accuracy of automated cephalometric analysis based on artificial intelligence. *BMC Oral Health*. 2023;23:191.
2. Pittayapat P, Limchaichana-Bolstad N, Willems G, Jacobs R. Three-dimensional cephalometric analysis in orthodontics: a systematic review. *Orthod Craniofac Res.* 2014;17:69–91.
3. Li C, Teixeira H, Tanna N, et al. The reliability of two- and three-dimensional cephalometric measurements: a CBCT study. *Diagnostics (Basel).* 2021;11.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-22 via free access

4. Hassan B, Nijkamp P, Verheij H, et al. Precision of identifying cephalometric landmarks with cone beam computed tomography in vivo. *Eur J Orthod*. 2013;35:38–44.
5. Yue W, Yin D, Li C, Wang G, Xu T. Automated 2-D cephalometric analysis on X-ray images by a model-based approach. *IEEE Trans Biomed Eng*. 2006;53:1615–1623.
6. Montufar J, Romero M, Scougall-Vilchis RJ. Hybrid approach for automatic cephalometric landmark annotation on cone-beam computed tomography volumes. *Am J Orthod Dentofacial Orthop.* 2018;154:140–150.
7. Shahidi S, Bahrampour E, Soltanimehr E, et al. The accuracy of a designed software for automated localization of craniofacial landmarks on CBCT images. *BMC Med Imaging.* 2014;14:32.
8. Gupta A, Kharbanda OP, Sardana V, Balachandran R, Sardana HK. A knowledge-based algorithm for automatic detection of cephalometric landmarks on CBCT images. *Int J Comput Assist Radiol Surg*. 2015;10:1737–1752.
9. Mohammad-Rahimi H, Nadimi M, Rohban MH, Shamsoddin E, Lee VY, Motamedian SR. Machine learning and orthodontics, current trends and the future opportunities: a scoping review. *Am J Orthod Dentofacial Orthop.* 2021;160:170–192 e174.
10. Hwang HW, Park JH, Moon JH, et al. Automated identification of cephalometric landmarks: Part 2-Might it be better than human? *Angle Orthod*. 2020;90:69–76.
11. Moon JH, Hwang HW, Yu Y, Kim MG, Donatelli RE, Lee SJ. How much deep learning is enough for automatic identification to be reliable? *Angle Orthod*. 2020;90:823–830.
12. Dot G, Rafflenbeul F, Arbotto M, Gajny L, Rouch P, Schouman T. Accuracy and reliability of automatic three-dimensional cephalometric landmarking. *Int J Oral Maxillofac Surg*. 2020; 49:1367–1378.
13. Gupta A, Kharbanda OP, Sardana V, Balachandran R, Sardana HK. Accuracy of 3D cephalometric measurements based on an automatic knowledge-based landmark detection algorithm. *Int J Comput Assist Radiol Surg.* 2016;11:1297–1309.
14. Ghowsi A, Hatcher D, Suh H, et al. Automated landmark identification on cone-beam computed tomography: accuracy and reliability. *Angle Orthod*. 2022;92:642–654.
15. Serafin M, Baldini B, Cabitza F, et al. Accuracy of automated 3D cephalometric landmarks by deep learning algorithms: systematic review and meta-analysis. *Radiol Med*. 2023;128: 544–555.
16. Gillot M, Miranda F, Baquero B, et al. Automatic landmark identification in cone-beam computed tomography. *Orthod Craniofac Res*. 2023;26:560–567.
17. Kielczykowski M, Kaminski K, Perkowski K, Zadurska M, Czochrowska E. Application of artificial intelligence (AI) in a cephalometric analysis: a narrative review. *Diagnostics (Basel)*. 2023;13.
18. Ahn J, Nguyen TP, Kim YJ, Kim T, Yoon J. Automated analysis of three-dimensional CBCT images taken in natural head position that combines facial profile processing and multiple deep-learning models. *Comput Methods Programs Biomed*. 2022;226:107123.
19. Hwang HW, Moon JH, Kim MG, Donatelli RE, Lee SJ. Evaluation of automated cephalometric analysis based on the latest deep learning method. *Angle Orthod*. 2021;91:329–335.
20. Lindner C, Wang CW, Huang CT, Li CH, Chang SW, Cootes TF. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. *Sci Rep*. 2016;6:33581.
21. Schlicher W, Nielsen I, Huang JC, Maki K, Hatcher DC, Miller AJ. Consistency and precision of landmark identification in three-dimensional cone beam computed tomography scans. *Eur J Orthod*. 2012;34:263–275.
22. Lee H, Cho JM, Ryu S, et al. Automatic identification of posteroanterior cephalometric landmarks using a novel deep learning algorithm: a comparative study with human experts. *Sci Rep*. 2023;13:15506.
23. Baumrind S, Frantz RC. The reliability of head film measurements. 2. Conventional angular and linear measures. *Am J Orthod*. 1971;60:505–517.
24. Baumrind S, Frantz RC. The reliability of head film measurements. 1. Landmark identification. *Am J Orthod*. 1971;60: 111–127.
25. Park J, Baumrind S, Curry S, Carlson SK, Boyd RL, Oh H. Reliability of 3D dental and skeletal landmarks on CBCT images. *Angle Orthod*. 2019;89:758–767.
26. Moon J-H, Lee J-M, Park J-A, Suh H, Lee S-J. Reliability statistics every orthodontist should know. *Semin Orthod*. 2024;30:45–49.
27. Moon JH, Shin HK, Lee JM, et al. Comparison of individualized facial growth prediction models based on the partial least squares and artificial intelligence. *Angle Orthod*. 2024; 94:207–215.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-06-22 via free access